THE TURKISH NARROW BAND VOICE CODING AND NOISE PRE-PROCESSING NATO CANDIDATE

Ahmet Kondoz *Hasan Palaz**

TÜBİTAK-UEKAE National Research Institute of Electronics & Cryptology
P.O. Box 21, 41470, Gebze, KOCAELI, TURKEY.
*E_mail : palaz@mam.gov.tr

ABSTRACT

Robust and low power communication systems are essential for battle field environment in military communication which require bit rates below 4.8kb/s. In order to benefit from the new advances in speech coding technologies and hence upgrade its communication systems, the NATO has been planning to select a speech coding algorithm with its noise pre-processor. In this paper we describe a speech coder which is capable of operating at both 2.4 and 1.2kb/s, and produce good quality synthesised speech. This coder will form the basis of the Turkish candidate which is one of the three competing. The rate of the coder can be switched from 2.4kb/s to 1.2kb/s by increasing the frame length for parameter quantisation from 20ms to 60ms. Both rates use the same analysis and synthesis building blocks over 20ms. Reliable pitch estimation and very elaborate voiced/unvoiced mixture determination algorithms render the coder robust to background noise. However in order to communicate in very severe noisy conditions a noise pre-processor has been integrated within the speech encoder.

1. INTRODUCTION

Speech coding at low bit rates has been a subject for intense research over the last 2 decades and as a result many speech coding algorithms have been standardised with bit rates ranging from 16kb/s down to 2.4kb/s. The standards covering the bit rates down to around 5kb/s are based mainly on CELP derivatives and the standards below 5kb/s are based mainly on frequency domain vocoding (harmonic coding) models such as sinusoidal coding [1]. Although in principle a harmonic coder should produce toll quality speech at around 4kb/s and good communications quality at around 2.4kb/s and below, various versions may have significantly different output speech quality. This quality difference comes from the way the parameters such as pitch and voicing are estimated/extracted at the analysis and the way parameters are interpolated for smooth evolution of the output speech during the synthesis process. A further difference is the parameter update rates and quantisation methods used. In this paper we focus on the split-band LPC (SB-LPC) approach to achieve a mode switchable 2.4-1.2kb/s coding rates with high intelligibility and good quality output speech, even during high background and channel noise conditions. Both versions of the algorithm work on 20ms analysis blocks and use the same analysis/synthesis procedures where a novel pitch detection algorithm and an elaborate voicing mixture determination are used which are essential for good speech quality. Although this algorithm performs well in background noise conditions, if the noise is too high (SNR<10dB) the use of a noise pre-processor (NPP) helps to improve the speech intelligibility as well as enabling perceptually more comfortable speech quality. We have therefore incorporated a NPP in the encoder.

In the following we present the description of the speech analysis/encoding, parameter quantisation followed by decoding/speech synthesis building blocks. This is then followed by the description of the NPP, and finally test results and the conclusions of the paper are presented.

2. SPEECH ANALYSIS

The Split-Band LPC Vocoder has been presented in detail in [2]. In this new version we have used a novel pitch estimation and a multiple input time/frequency domain voicing mixture classification algorithms. Residual spectral magnitudes are extracted by selecting the harmonic peaks for the voiced part of the spectrum and computing the average noise energy in each fundamental frequency band for the unvoiced part. During the extraction of the residual spectral magnitudes we are only interested in the relative variations of magnitudes and not their absolute values. A separate energy control factor is computed from the input speech for proper scaling of the signal at the output of the synthesiser. Speech analysis and synthesis are based on 20ms frames but parameters are quantised every 20ms for 2.4kb/s and every 60ms for 1.2kb/s versions respectively.

2.1 PITCH ESTIMATION ALGORITHM

The pitch estimation algorithm consists of three parts. First a frequency domain analysis is performed. The most promising candidates from this first search are then checked by computing a time domain metric for each. Finally one of the remaining candidates is selected based on the frequency and time domain metrics, as well as the tracking parameters.

Frequency domain pitch analysis is performed using a modified version of the algorithm described by McAulay [4] which determines the pitch period to half sample accuracy. The speech is windowed using a 241 point Kaiser window ($\beta = 6.0$), then a 512 point FFT is performed to obtain the speech spectrum. The fundamental frequency is the one that produces the best periodic fit to the smoothed spectrum. In order to reduce complexity, only the lower 1.5 kHz of this spectrum is used for the pitch

Paper presented at the RTO IST Symposium on "New Information Processing Techniques for Military Systems", held in Istanbul, Turkey, 9-11 October 2000, and published in RTO MP-049.

algorithm. To further reduce complexity, only integer pitch values are used above the pitch value of 45 samples.

However, this initial pitch estimate is not always correct. In particular doubling and halving of the pitch frequency can occur. In order to avoid these problems, a certain number of candidate pitch values are selected for further processing. In addition, the range of possible values for ω_0 is divided into 5, corresponding in pitch lags of: [15-27],[27.5-49.5],[50-94.5],[95-124.5] and [124.5-150]. In each of these intervals, the best candidate is also selected, if it is not already selected in the first stage. These intervals are selected so that no pitch candidate can double in a given interval.

All candidate pitch periods determined above are re-examined using a metric which measure the RMS energy variations with respect to the energy computation block length which takes the values given by the candidate pitch periods. The RMS energy fluctuation is minimum when RMS computation block length equals the correct pitch period or its integer multiples.

After the elimination of some candidates based on the time domain metric, if more than one pitch candidates are left, the final decision process operates as follows: For each candidate a final metric is computed, which takes into account both the time- and frequency- domain measures: The candidate with the best combined final metric is then selected as a pitch estimate. In order to avoid pitch doubling, a sub-multiple search is performed. If there is a remaining candidate close enough to being a sub-multiple of the current pitch estimate, and whose final metric is above a certain threshold (typically 0.8 times the final metric of the current pitch estimate), then it is selected as the new current pitch estimate. The sub-multiple search is then repeated using this new value.

The pitch algorithm described above is usually reliable in clean speech conditions. However, it occasionally suffers from pitch doubling and halving when the pitch is not clearly defined, or in heavy background noise conditions. To overcome this problem we have used a mild pitch tracking. In order to be able to update the tracked pitch parameters during speech only frames a simple voice activity detector which is explained in section 5 is used. After the computation of the time and frequency domain metrics, before the start of the elimination process, each candidate which is close to the tracked pitch has its metrics biased to increase its chances of being selected as the final pitch.

The VAD also determines the signal to background noise ratio of the input samples which controls the amount of tracking used. The bias applied by tracked pitch on the metrics is more for noisy speech than in clean speech conditions.

In clean speech conditions this pitch estimation algorithm exhibits very few errors. They only occur when the pitch is not clearly defined and only extra look-ahead could improve this. It is also very resilient to background noise, and still operates satisfactorily down to SNR of 5 dB. At higher noise levels errors start to occur occasionally but the algorithm still manages to give the correct pitch value most of the time.

2.2 LP EXCITATION VOICING MIXTURE

Many low bit rate vocoders now use the assumption that the voicing content of the speech can be represented by only one cut-off frequency below which the speech is considered harmonic and above which it is considered stochastic. This has the advantage of requiring only a very small number of bits to quantise the voicing information, as opposed to transmitting one bit per harmonic band. If performed accurately, the distortion induced by this assumption will be very limited and acceptable for low bit rate speech coders. It is however very important to correctly determine the cut-off frequency as errors will induce large distortions in the output speech quality.

In SB-LPC, for accurate voicing extraction the speech is first windowed using a variable length Kaiser window. Four different windows are used, from 121 to 201 samples in length, depending on the current pitch period, so as to have the smallest possible window covering at least 2 pitch cycles. In the next step the limits of each harmonic band across the spectrum is determined. This is done by refining the original pitch estimate down to a more accurate fractional pitch. The original pitch accuracy is at half a sample accuracy up to the pitch value of 45 samples and integer for bigger values. Moreover the pitch has been determined using only the lower 1.5 kHz of the spectrum. The spacing of the harmonics might be slightly different in the higher part of the spectrum. Hence it is necessary to refine the pitch using the whole of the 4 kHz spectrum.

A threshold value is then computed for each band across the spectrum, based on various time- and frequency domain factors. The general idea being that if the voicing value is above the threshold value for a given band, then it is probably voiced. Finally for each possible quantised cut-off frequency, a matching measure is computed using the threshold and voicing measures for each band, and the final quantised cut-off frequency is selected as the one which maximises this matching.

If a harmonic band is voiced, then its content will have a shape similar to the spectral shape of the window used to window the original speech prior to the Fourier transform, whereas unvoiced bands will be random in nature. Hence voicing can be determined by measuring the level of normalised correlation between the content of the harmonic band and the spectral shape of the window. The normalized correlation lies between 0.0 and 1.0, where 0.0 and 1.0 indicates unvoiced and voiced extremes respectively.

For the decision making this normalized correlation is compared against a fixed threshold for each band across the spectrum. Since the likelihood of voiced and unvoiced is not fixed across the frequency spectrum, and may also vary from one frame to the next, the decision threshold value needs to be adaptive for accurate voicing determination. When determining a voicing threshold value for each frequency band (harmonic) we have used additional factors some of which are listed in [3]. A threshold value is computed for each band based on the following variables:

- the peakiness (ratio of the L1 to L2 norms),
- the cross-correlation value at the pitch delay,

- the ratio of the energy of the high frequencies to energy of the low frequencies in the spectrum
- the ratio between the energies of the speech and of the LP residual
- the ratio between the energy of the frame and the tracked maximum energy of the speech, E_s/E_{max}.
- the voicing of the previous frame
- a bias is added to tilt the threshold toward more voiced in the low frequencies.

Having computed a voicing measure and a threshold for each harmonic band we now need to find the best quantised cut-off frequency for this set of parameters. For each possible quantiser value a matching measure is computed taking into account the difference between the correlation value and the corresponding threshold, as well as the energy in a given harmonic band. A bias which favors voiced decisions over unvoiced decisions is also used. A typical quantiser for the voicing is a 3 bits quantiser, representing 8 cut-off frequencies spaced between 0 and 4 kHz.

3. PARAMETER QUANTISATION

Table 1. shows the bit allocation for the 2.4 and 1.2kb/s versions.

Bit Rate	2.4 kb/s	1.2 kb/s		
Update rate (in ms)	20	60		
LPC	21	44		
Pitch	7	3	6	3
Voicing	3	3		
RMS energy	6+1	6+6		
Spectral Magnitudes	9	0	0	0
Sync. bit	1	1		
Total	48	72		

Table 1: Bit allocation for the different rates of the Split-Band LPC Vocoder

In the case of 2.4kb/s 47 bits are used to quantised the parameters every 20ms. The LP parameters are quantised in the form of line spectral frequencies (LSF) with a multi-stage vector quantisation (MSVQ) which has three stages of 7,7,7 bits. However, before the MSVQ, a first order moving average (MA) prediction with 0.5 predictor is applied to remove some of the correlation in the adjacent LP parameter sets. The RMS frame energy is quantised with a 6-bit scalar quantiser after a similar MA prediction with 0.7 predictor plus one bit protection. Only the 64 levels out of the 128 (6+1 bits) are used for encoding by ensuring that in case of channel errors, the codewords that could potentially result in large gain changes are not used. This process ensures that the errors introduced will have minimum damaging effect. The pitch is quantised non-uniformly with 7-bits, covering the range from 16 to 150 samples. Since the residual spectral magnitudes under the formant regions are more important, during magnitude quantisation the most important 7 magnitudes followed by the average value of the rest is vector quantised using a 9-bit codebook.

In the case of 1.2kb/s, a frame of 60ms is used where it is split into three 20ms sub-frames. The LP parameters are multi stage vector quantised using 44bits after a similar MA prediction process. For the pitch, voicing and energy computations, 20ms sub-frame length is used and repeated 3 times per frame. Pitch of the first and third sub-frames are quantised with respect to the pitch of the middle sub-frame using 3-bits each. The middle sub-frame's pitch is quantised using 6-bits. The voicing mixtures of all three sub-frames are jointly quantised using 3-bits. Similarly the RMS energies are jointly quantised with a gain shape vector quantiser using 6 bits for the gain and 6 bits for the three element shape vector.

4. DECODING AND SPEECH SYNTHESIS

4.1 Parameter Decoding

In the 2.4kb/s mode, each 20ms frame has its own LP parameters, pitch, voicing mixture and the RMS frame energy which are sufficient for good quality speech synthesis. During the decoding process of LSFs the usual stability checks are applied. When decoding the RMS energy, channel error effects are minimised by using only 64 possible combinations of the 7 bits representation with proper robust index assignment [5]. For the pitch and voicing no channel error checks are applied.

In the case of 1.2kb/s no error checks are applied to any of the parameters, except the usual LSF stability check and robust index assignment [5].

4.2 Speech Synthesis

In order to improve the speech quality, at the decoder we introduce half a frame delay for both 2.4 and 1.2kb/s versions. In the case of 2.4kb/s first half of 20ms frame is synthesised by interpolating the current parameters with the preceding set and the second half uses the parameters interpolated between the current and the next sets. Similar interpolation is applied for the 1.2kb/s version where each 20ms sub-frame is assumed to be a 20ms frame. The actual interpolation is applied pitch synchronously and the contribution of the left and right hand side parameters is based on the centre position of each pitch cycle within the synthesis frame. The actual synthesis of both voiced and unvoiced sounds is performed using an IDFT with pitch period size. The voiced part of the spectrum has only the magnitudes with zero phases and the unvoiced part of the spectrum is filled with both unvoiced magnitudes and random phases. If desired a perceptual enhancement process is applied where the valley regions of the excitation spectrum are suppressed [2]. The resultant excitation is then passed through the LP synthesis filter which has its parameters interpolated pitch synchronously. Finally the output signal which may have arbitrary energy is normalised per pitch cycle to match the interpolated frame energy.

5. NOISE PRE-PROCESSOR

The SB-LPC speech coder with the above detailed parameter analysis and quantisation techniques operate well within background noise environments. However, both speech quality and intelligibility in heavy noise conditions can be improved if a suitable noise suppression/pre-processing technique (NPP) is used before speech analysis is applied. We have used a noise pre-processing technique to suppress the background noise before encoding [8][9]. A significant reduction of the background noise level improved the parameter estimation process which improved the overall synthesized speech quality in the presence of noise. Furthermore reduction of the overall noise enables a more comfortable listening level which is very significant in terms of the tiredness it may cause to the user. The performance of the NPP is dependent on the speed of adaptation of its parameters and correct voice activity detection (VAD). The VAD used in [8] compares the ratio of the current frame's power and the accumulated noise power against a pre-set threshold which works well in reasonably high SNR conditions (typically 10dB or greater). When the SNR worsens this VAD makes occasional mistakes in declaring noise as speech mixed with noise, and speech mixed with noise as noise only. The former reduces the speed of adaptation of the background noise which is not very serious. The latter on the other hand updates background noise while speech present which causes significant distortion in the output speech quality.

We have used an energy-dependent time-domain VAD technique, which helps in better tracking speech and noise levels during harsh background noise conditions. This VAD algorithm estimates the levels of various energy parameters - instantaneous energy E_0, minimum energy E_{min}, maximum energy E_{max} - that are, in turn, used to indicate the SNR estimate of the current frame. The role of E_{max} is to track the maximum value of the input signal, which is done by a slow descending and sharp ascending adaptation characteristic. E_{min} tracks the minimum energy of the input signal and is therefore characterised by a sharp descending and slow ascending gradient. The SNR_{est} represents the ratio between the maximum and the minimum energy for any given frame.

The importance of the SNR_{est} is that its level controls the energy thresholds for the VAD. Namely, the VAD operates according to the ratio:

$$VAD=\begin{cases} 0, (E_0/E_{min})<E_{th} \\ 1, (E_0/E_{min})\geq E_{th} \end{cases}$$

where the value of E_{th} depends on the SNR estimate and is adaptively constrained to be within a limited range of 1.25-2.0.

Another important feature of the SNR_{est} is that it defines the speed of adaptation for the NPP parameters.

In order to reduce the overall NPP+speech encoding/decoding delay, the NPP frame size (up-date rate) must be same as or integer sub-multiple of the speech frame. The NPPs usually have 256 sample window and FFT building blocks which are shifted by 128 samples (up-date rate). A Hanning window is usually preferred since the synthesis process becomes a simple overlap and add. However the up-date rate of 128 samples is unsuitable for the 20ms speech frames. We have therefore used 80 samples up-date rate (176 samples overlap) and applied two NPP processes per speech frame. Since the overlap of the two adjacent NPP processing stages is more than 50%, during the NPP cleaned speech synthesis the two adjacent blocks are first de-windowed (to remove the analysis windowing effect) and then a trapezoidal window is used before overlap/add is executed.

6. SIMULATIONS

In order to assess the performance of the designed coder we have used subjective listening tests. In the tests 2 male and 2 female speakers with two sentences from each were used. The input sentences were also added with noise at 10 and 5dB. Three types of noise were used, helicopter, vehicle and bable. The input level of the signal was set to nominal -26dB during all testing. In the tests A and B comparisons were made. Each sentence was played twice one produced by our coder and one produced by the reference coder. We have used two reference coders, the DoD CELP at 4.8kb/s [6] and MELP at 2.4kb/s [7]. During the comparisons 22 trained subjects were asked to grade their preferences using 2, 1, 0, -1 and -2 to indicate better, slightly better, the same, slightly worse and worse respectively. They were also asked to describe the reasons for their choice.

The coders were numbered as C1, C2 and C3 for SB-LPC at 2.4kb/s, 1.2kb/s and 2.4kb/s+NPP respectively. The reference coders were numbered as R1 and R2 for CELP and MELP respectively.

Comparison	Clean Speech	Noisy Speech
C1 vs. R1	11	-2
C1 vs. R2	9	13
C1 vs. C2	2	1
C1 vs. C3	0	-10

Table 2: Subjective comparison results

As can seen from the results in Table 2, in clean speech there is a clear preference for SB-LPC as compared with DoD CELP. The main reason for not preferring CELP was its rather noisier output quality. The quality of the SB-LPC has been preferred due its cleanness and less muffling. In noisy speech however the preference of CELP was found to increase. There were two main reasons for this. Firstly the reproduction of the background noise by CELP had a more pleasant nature and it was easier to recognize the noise type. The second reason is that since the voicing classification of the SB-LPC was tuned to favor voiced, during the noise only parts some voiced declarations caused periodic components which were found to be unpleasant.

When compared against MELP under clean background conditions SB-LPC was preferred again. The main reason for this was that MELP had occasional artifacts which was found to be annoying and had more metallic nature. Under background noisy conditions the difference was more noticeable. The reason for this difference was that MELP voicing decision mistakes caused roughness in its output speech quality. Some on-sets and off-sets

where the relative noise level was high, were declared as unvoiced.

After the comparison of the 2.4kb/s SB-LPC against the two DoD standards it was then compared against its 1.2kb/s version. In clean speech input case, there was a slight preference for the 2.4kb/s. In the noisy conditions, as expected, the two rates were found to be very similar. The comparison of the 2.4kb/s with and without NPP clearly showed the NPP's effectiveness in noisy conditions. Finally the 2.4kb/s version was informally tested under 1% random bit errors and 3% frame erasures. Although, the random bit errors caused slight degradations, owing to accurate frame substitution methods, frame erasures did not caused noticeable distortions.

7. CONCLUSIONS

In this paper we have presented a split-band LPC based speech coder which is capable of operating at two modes of 2.4 and 1.2kb/s. Both of the modes use the same core analysis and synthesis blocks. The rate halving is obtained by increasing the encoding delay to have efficient quantisation of the parameters with fewer bits. A noise pre-processor has also been integrated with the speech encoder to improve the performance during noisy background conditions.

The coder was tested in two stages. In the first stage the 2.4kb/s version was compared against DoD CELP and MELP algorithms operating at 4.8 and 2.4kb/s respectively. In the second stage two modes of the coder were compared to quantify the degradation incurred in halving the bit rate. In clean input condition the 2.4kb/s version was preferred against both references but in noisy speech condition CELP was found to be slightly better. In the case of 1.2kb/s very similar speech quality to the 2.4kb/s version was produced for both clean and noisy inputs. The use of a NPP at the encoder increased the performance of the coder for noisy input samples. Both speech intelligibility and quality was improved significantly. The 2.4kb/s version was also tested against channel errors at 1% random bit error rates and 3% frame erasure rates. The random bit errors were found to cause slight quality reductions. However by protecting the RMS energy with a single bit possible blasts were eliminated. The 3% frame erasures did not cause noticeable degradation.

8. REFERENCES

[1] R.J. McAulay, T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. on ASSP*, 34 pp 744-754, 1986.

[2] I.Atkinson, S.Yeldener, A.M.Kondoz, "High Quality Split-Band LPC Vocoder Operating at Low Bit Rates" *ICCASP-97*, Volume 2, pp 1559-1562.

[3] J.P Campbell, T.E Tremain "Voiced/unvoiced classification of speech with applications to the U.S. Government LPC-10E Algorithm", *ICASSP -1986, pp 9.11.1-9.11.4.*

[4] R.J. McAulay, T. F. Quateri, "Pitch Estimation and Voicing Decision Based Upon A Sinusoidal Speech Model", *ICASSP-90*, Vol. 1, pp 249-252.

[5] K.Zeger, A.Gersho, "Pseudo-Gray Coding", *IEEE Trans. On Communications*, 38, no 12, pp 2147-2156, 1990.

[6] J. P. Campbell, T. Tremain, V. C. Welsh, "The DoD 4.8kbps Standard (Proposed Federal Standard 1016)", *Speech Technology*, Vol. 1(2), pp 58-60, April 1990.

[7] A. McCree et.al. "A 2.4kb/s MELP Coder Candidate for the New U.S. Federal Standard", *ICASSP-96*, pp 200-203.

[8] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. ASSP-32, No.6 pp. 1109-1121, December 1984.

[9] R.J. McAulay, M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-28, NO.2, April 1980, pp 137-145.

This page has been deliberately left blank

Page intentionnellement blanche